DOCUMENT RESUME

ED 400 270 TM 025 388

AUTHOR Rachor, Robert E.; Gray, George T.
TITLE Must All Stems Be Green? A Study of Two Guidelines for Writing Multiple Choice Stems.
PUB DATE Apr 96
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Allied Health Personnel; *Difficulty Level; High Achievement; Item Banks; *Licensing Examinations (Professions); Low Achievement; *Multiple Choice Tests; Physicians; Reading Comprehension; *Test Construction; Test Format; Test Items
IDENTIFIERS American College Testing Program; Item Discrimination (Tests); T Test

ABSTRACT

Two frequently cited guidelines for writing multiple choice test item stems are: (1) the stem can be written in either a question or statement-to-be-completed format; and (2) only positively worded stems should be used. These guidelines were evaluated in a survey of the test item banks of 13 nationally administered examinations in the physician specialty and allied health professions prepared by the American College Testing program. These items had been written and reviewed by content specialists, and had been professionally edited. Mean item difficulty and discrimination were compared for each test using separate t-tests. Although a relatively strong case for the question format and for positively worded item stems may be made based on considerations related to grammar and reading comprehension, this study offered little research support for these item writing rules. It must be recognized that these examinations have been prepared for subjects at high educational levels; item stem characteristics may be more important for lower achievement levels and less carefully edited tests. (Contains four tables and five references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
ROBERT E. RACHOR

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 270

Must All Stems Be Green? A Study of Two Guidelines For Writing Multiple Choice Stems

Robert E. Rachor, Ph.D.
Toledo Public Schools
420 E. Manhattan Blvd.
Toledo, OH 43608
(419) 729-8434

George T. Gray, Ed.D.
American College Testing
PO Box 168
Iowa City, IO 52243
(319) 337-1168

Paper Presented at the Annual Meeting of the
American Educational Research Association
New York City
April, 1996

TM025388

BEST COPY AVAILABLE

Introduction and Review of Literature

Specific multiple choice item writing rules have been widely promoted for many years, even though many of the rules have not been extensively researched. Two frequently cited guidelines for writing multiple choice item stems are: (1) the stem can be written in either a question or statement to be completed format, and (2) use only positively worded stems.

Haladyna and Downing (1989a) reported that 41 out of 46 "authoritative references in the measurement literature" supported the use of either the question or completion format when writing the stem of an item. They noted however, that 6 research studies reported a mean decrease in item difficulty using the question format. Haladyna and Downing (1989b) suggested revising the rule to read, "use the question format, avoid the completion format" (p. 71). Haladyna (1994) took a firm position on this issue, suggesting that, "From the standpoint of reading comprehension, administration time, anxiety, the quality of the test item, and efficiency, the question format is a more effective way to test for knowledge than the completion format" (p.70).

Haladyna and Downing (1989a) reported that 31 out of their 46 "authoritative references" recommended wording the stem of the item positively, while four references indicated that either positive or negative stems were appropriate. Haladyna and Downing (1989b) concluded that, "the rule (use positive stems only) appears valid pending more definitive research." Recently Harasym et al (1992, 1993) reported studies in which two forms of an examination were administered to 200 nursing students. These forms contained positive and negative versions of ten items. In order to make positive and negative versions of the items equivalent, the items were transformed from single-response items having a negatively worded stem to multiple response items having a positive stem. The multiple response scoring was achieved by awarding fractional credit for a correct response to each alternative. These studies suggested that negatively worded stems lowered item reliability and reduced item difficulty by inadvertently cueing the correct answer and argued for a multiple-response scoring system with positively worded stems.

Published studies of multiple choice item performance characteristics have used a variety of methods. Experimental designs using a small number of items have been the norm. Experimental

designs have been better suited to the investigation of items having stems in question or completion format because the stem format can be interchanged without changing the options or scoring rubric. Post hoc studies, that have not controlled for item content, have also been reported. A number of variables have been investigated including difficulty and discrimination indices. Oddly enough, difficulty and discrimination have been studied as separate variables, although a multiple choice item with desirable statistics is characterized by the combination of its difficulty and discrimination. (A desirable item usually has difficulty that falls in an intermediate range and some discriminating power. Items that have less desirable characteristics on one of these dimensions may have optimum difficulty and no discrimination or have discriminating power but be extremely easy or hard.) The current study attempted to address this latter problem by combining difficulty and discrimination indices to categorize items as either desirable or undesirable.

Methodology

The present study was based on a survey of item banks used for 13 nationally administered examinations in the physician specialty and allied health professions. These examinations were either high stakes certification or self-assessment examinations. Items used were either embedded pretest items or newly written items that were included in the scored portion of the exam. All embedded pretest items or newly written items in each of the 13 nationally administrated examinations were used. Because all examination items were new, their inclusion in an examination was based strictly on content considerations and not on their statistical performance. These items had been written and reviewed by content specialists and professionally edited. Some examination committees have particular editorial preferences (such as limiting the number of items on the examination with negatively worded stems); however, the authors are not aware of any aspect of the item writing and review process that would be expected to have a systematic influence on the performance of items with particular editorial characteristics.

The post hoc approach was used because of the large numbers of available items from different sources; however it was somewhat problematic because item content was not controlled

with other attributes of the item that might be expected to affect item performance. To address this limitation an additional study was conducted to determine if the impact of content sampling affected performance on item difficulty and discrimination.

Sampling of Items

Items were studied from item banks of clients served by the Health Programs Department of American College Testing. Because the characteristics of examinees differed by the specialty of the health profession, the data were analyzed separately by client group. In effect, there were a number of replications of the study for each guideline being investigated. Different examination administrations for a particular program were combined. The median number of examinees per form for the study on question and completion stems was 195; the average number of questions per client group was 455 (s=296). The median number of examinees per form for the study on positive and negative stems was 287; the average number of questions per client group was 328 (S=171).

Data Analysis

Prior to the study, a decision was made to not consider a specific examination for one of the two item writing guidelines studies unless at least 10% of the items met each of the two characteristics (i.e. positive or negative stems) of the particular guideline under study. In order to meet the assumptions of the parametric tests employed, item difficulty and discrimination indices were transformed using arcsin and Fisher's z respectively. The mean item difficulty and discrimination for sets of items for each guideline were compared separately using independent t-tests within each client group. The power of each parametric test to find differences that may have existed was also calculated.

Difficulty and discrimination indices were combined to designate desirable and undesirable items. Desirable performance characteristics for each item were arbitrarily defined as 40%-90% correct for item difficulty and .10 or greater for item discrimination (point biserial correlation). A 2x2 chi-square analysis was performed based on the item characteristic (e.g., positively or negatively worded stems) and whether or not the items met both desirable performance characteristics.

Results

A. Question vs. completion format

Table 1 presents descriptive and inferential statistics for the investigation of the 13 programs on the performance of sentence and completion items. Mean item difficulty and discrimination indices were compared for each client program using separate t-tests. Lower item difficulty indices ($p \leq .05$) were found for the question format on two of the thirteen programs. There were significant differences in item discriminations indices ($p \leq .05$) for two other programs; in both instances the question format had a higher item discrimination indice than the completion format. Separate bivariate plots of difficulty and discrimination using the question and completion format were remarkably similar in shape and density for most programs.

Table 1 About Here

The power of the univariate t-tests to find differences that may have existed in item difficulty and discrimination for each of the 13 examinations was also calculated. The power of the univariate tests to detect differences in item difficulty ranged from .72 to .98 with only eleven of the thirteen examinations having power of at least .80. The power to detect differences in item discrimination ranged from .59 to .93 with seven of the thirteen examinations having power of at least .80.

Table 2 About Here

Table 2 presents differences between desirable and undesirable items for question or completion stems. A chi-square test of proportions was used to test whether the differences were significant. Significant differences in the proportion of desirable and undesirable items ($p \leq .05$) using the question or completion formats were found in only one examination. The one program with significant results had a greater proportion of question stems with desirable indices of difficulty

and discrimination than completion items. Results of the previous analysis for the same client program found that question stems were more discriminating than completion stems.

Because item content was not controlled in this study, additional data analysis was performed to obtain some indication of the influence of content sampling on the results of the study. For each guideline studied, the two formats (question vs. completion stems) were randomly assigned to two groups. These paired samples revealed no differences in difficulty, discrimination, or proportion of items with preferred statistical attributes.

B. Positive vs. negative wording of stem

Because many examination committees avoid negatively worded stems, only seven client programs were available for this portion of the study. Table 3 summarizes descriptive and inferential statistics for the seven programs using positive and negative stems. In all programs, the negative word in the stem was capitalized and in bold print. Positive items were easier than negative items in one program ($p \leq .01$). There were no differences in item discrimination on any of the seven examinations.

The power of the univariate t-tests to find differences that may have existed in item difficulty and discrimination for each of the seven examinations was also calculated. The power of the univariate tests to detect differences in item difficulty ranged from .57 to .85 with only two of the seven examinations having power of at least .80. The power to detect differences in item discrimination ranged from .56 to .87 with only three of the seven examinations having power of at least .80.

Table 3 About Here

Table 4 presents differences between desirable and undesirable items for positive and negative stems. A chi-square test of proportions was used to test whether the differences were significant. The one examination with significant differences in difficulty also exhibited significant differences ($p \leq .05$) in the proportion of desirable and undesirable items between positive and

negative question stems. While the results for that examination concluded that positive items were less difficult, a greater proportion of negative items had more desirable characteristics. There were no differences in the proportion of desirable or undesirable items between the positive and negative formats in the other six programs. Again, separate bivariate plots of difficulty and discrimination using the question and completion format were remarkably similar in shape and density.

Table 4 About Here

Because item content was not controlled in this study, additional data analysis was performed to obtain some indication of the influence of content sampling on the results of the study. For each guideline studied, the two formats (positive and negative stems) were randomly assigned to two groups. These paired samples revealed no differences in difficulty, discrimination, or proportion of items with preferred statistical attributes.

Discussion

Item writing guidelines may be supported by both logic and research. A relatively strong case for the question format and positively worded stems may be made based on considerations related to grammar and reading comprehension; however, this study offered little research support for these item writing rules.

There are a number of possible reasons for these findings. First, all but one of the examinations included in this study required at least a bachelor's degree for entry level; many required an advanced degree. It may be that for higher educational levels, these two characteristics of the item stem have little influence on performance compared to the specialized knowledge required for answering items in the content areas.

Second, peer review by examination committees and the process of editing items for clarity may minimize any adverse impact caused by the form of the item stem. Item editing is a variable that has received little attention in the literature. Most studies have not included items that were

BEST COPY AVAILABLE

professionally edited. Third, there may not have been enough power to detect differences between positive and negative stems.

Finally, the results might be attributed to the limitations of the study methodology. As previously indicated, item content was not experimentally controlled. Despite this limitation, the study includes a number of nationally administered examinations rather than a single examination administered to a local population. In addition, there is no particular reason to suspect that the sampling of content would systematically reflect differences in item difficulty or discrimination for one stem format or another, given the fact that the first-time administration of all available items is included in the study.

The preference for a question or completion stem would seem to rest primarily on logical grounds. This guideline is more amenable to research than a number of others, and additional empirical support (or lack thereof) for the guideline should be obtained. Haladyna's argument for the question format based on reading comprehension, administration time, anxiety, quality of the test item and efficiency seems to be overstated based on data that is presently available.

We should be aware of the fact that most item writing guidelines are only loosely grounded in research and that the little research that has been done is not usually based on samples of students in professional programs or candidates for certification or licensure in a profession. There are good reasons for avoiding negatively worded stems in multiple choice items. Anyone who has puzzled over the pairing of a negatively worded stem and a negatively worded alternative probably needs no persuasion concerning this guideline; however, there may be occasions in which it is desirable to evaluate an individual's knowledge of an exception. The negatively worded stem may, given appropriate editorial precautions, be well-suited to such a task. This research suggests that the negative wording of a stem does not necessarily pose a hurdle for health professionals that is large enough to be measured.

Table 1

Differences Between Question and Completion Stems

Exam	Examinees	# Items Question or Completion	Item Difficulty or Discrimination	Mean Question	Mean Completion	T-test Power	T-test Probability
1	403	130/159	Item Difficulty	0.65	0.63	0.72	0.42
			Discrimination	0.19	0.18	0.70	0.34
2	595	87/113	Item Difficulty	0.60	0.63	0.92	0.23
			Discrimination	0.25	0.24	0.74	0.43
3	153	128\152	Item Difficulty	0.67	0.70	0.93	0.36
			Discrimination	0.17	0.17	0.83	0.96
4	287	84/135	Item Difficulty	0.69	0.72	0.95	0.21
			Discrimination	0.17	0.16	0.77	0.60
5	1714	181/215	Item Difficulty	0.61	0.63	0.90	0.44
			Discrimination	0.19	0.20	0.90	0.58
6	5783	381/791	Item Difficulty	0.56	0.59	0.98	0.03
			Discrimination	0.20	0.20	0.82	0.84
7	177	129/111	Item Difficulty	0.63	0.65	0.90	0.54
			Discrimination	0.25	0.26	0.92	0.55
8	529	130/145	Item Difficulty	0.66	0.67	0.87	0.80
			Discrimination	0.22	0.20	0.67	0.27
9	193	282/443	Item Difficulty	0.57	0.61	0.98	0.05
			Discrimination	0.14	0.15	0.93	0.33
10	113	124/596	Item Difficulty	0.62	0.61	0.86	0.84
			Discrimination	0.16	0.15	0.75	0.55
11	109	146/334	Item Difficulty	0.74	0.73	0.79	0.69
			Discrimination	0.22	0.18	0.59	0.02
12	195	251/469	Item Difficulty	0.73	0.75	0.96	0.13
			Discrimination	0.17	0.13	0.92	0.00
13	183	51/149	Item Difficulty	0.71	0.74	0.89	0.63
			Discrimination	0.14	0.14	0.85	0.96

Table 2

Differences Between Desirable/Undesirable Items for Question or Completion Stems

Exam	Examinees	Question/ Completion	Desirable Undesirable Items	Question Stems	Completion Stems	Chi-Square p
1	403	130/159	Desirable	66	92	0.228
			Undesirable	64	67	
2	595	87/113	Desirable	63	84	0.760
			Undesirable	24	29	
3	153	128\152	Desirable	66	86	0.401
			Undesirable	62	66	
4	287	84/135	Desirable	50	70	0.267
			Undesirable	34	65	
5	1714	181/215	Desirable	120	137	0.592
			Undesirable	61	78	
6	5783	381/791	Desirable	239	512	0.504
			Undesirable	142	279	
7	177	129/111	Desirable	74	73	0.183
			Undesirable	55	38	
8	529	130/145	Desirable	77	93	0.403
			Undesirable	53	52	
9	193	282/443	Desirable	127	221	0.202
			Undesirable	155	222	
10	113	124/596	Desirable	62	257	0.161
			Undesirable	62	339	
11	109	146/334	Desirable	76	178	0.802
			Undesirable	70	156	
12	195	251/469	Desirable	135	187	0.000
			Undesirable	116	282	
13	183	51/149	Desirable	17	65	0.197
			Undesirable	34	84	

Table 3

Differences Between Positive and Negative Stems

Exam	Examinees	# Items Positive or Negative	Item Difficulty or Discrimination	Mean Positive	Mean Negative	T-test Power	T-test Probability
1	403	202/87	Item Difficulty	0.65	0.60	0.57	0.01
			Discrimination	0.18	0.19	0.87	0.77
2	595	174/26	Item Difficulty	0.62	0.60	0.75	0.54
			Discrimination	0.25	0.24	0.80	0.77
3	153	242/38	Item Difficulty	0.68	0.67	0.82	0.89
			Discrimination	0.18	0.14	0.56	0.09
4	287	184/35	Item Difficulty	0.71	0.70	0.75	0.50
			Discrimination	0.17	0.14	0.63	0.19
5	1714	289/107	Item Difficulty	0.62	0.62	0.85	0.97
			Discrimination	0.20	0.19	0.76	0.59
6	193	587/138	Item Difficulty	0.60	0.58	0.69	0.32
			Discrimination	0.14	0.14	0.85	0.96
7	177	155/85	Item Difficulty	0.64	0.63	0.79	0.73
			Discrimination	0.26	0.24	0.65	0.24

Table 4

Differences Between Desirable/Undesirable Items for Positive or Negative Stems

Exam	Examinees	Positive/ Negative	Desirable Undesirable Items	Positive Stems	Negative Stems	Chi-Square p
1	403	202/87	Desirable	99	59	0.003
			Undesirable	103	28	
2	595	174/26	Desirable	127	20	0.672
			Undesirable	47	6	
3	153	242/38	Desirable	133	19	0.568
			Undesirable	109	19	
4	287	184/35	Desirable	101	19	0.947
			Undesirable	83	16	
5	1714	289/107	Desirable	184	73	0.399
			Undesirable	105	34	
6	193	587/138	Desirable	282	66	0.964
			Undesirable	305	72	
7	177	155/85	Desirable	94	53	0.795
			Undesirable	61	32	

References

Haladyna, T.M. Developing and Validating Multiple-Choice Test Items. Lawrence Erlbarm Associates. Hillsdale, NJ. 1994. 221 pgs.

Haladyna, T.M., and Downing, S.M. (1989a). A taxonomy of multiple-choice item writing-rules. Applied Measurement in Education, 2(1), 37-50.

Haladyna, T.M., and Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2(1), 51-78.

Harasym, P.H., Price, P.G., Brant, R., Violato, C., and Lorscheider, F.L. (1992). Evaluation of negation in stems of multiple-choice items. Evaluation & the Health Professions, 15(2), 198-220.

Harasym, P.H., Doran, M.L., Brant, R., and Lorscheider, F.L. (1993). Negation in stems of sing-response multiple-choice items. Evaluation & the Health Professions, 16(3), 342-357.